UNSUPERVISED DOMAIN ADAPTATION WITH MULTIPLE ACOUSTIC MODELS

Xin Lei[1,2,†] Wen Wang[1] Andreas Stolcke[1]

[1]SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025 USA
[2]Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043 USA
xinlei@google.com, {wwang,stolcke}@speech.sri.com

ABSTRACT

We investigate the problem of adapting a recognition system with multiple acoustic models to a new domain in unsupervised mode. We compare maximum likelihood and discriminative approaches for unsupervised domain adaptation. Different adaptation data selection methods and adaptation strategies are investigated, using a baseline meeting recognition system and adaptation data from a congressional committee web site. Experiments show that one should avoid adapting all acoustic models to the same recognition output, and that ASR confidence estimates improve results when used for rejecting low-quality ASR output. The results show 8% relative overall improvement from unsupervised adaptation.

Index Terms— domain adaptation, unsupervised adaptation, discriminative adaptation

1. INTRODUCTION

We are concerned with unsupervised adaptation of acoustic models for state-of-the-art, multipass, large-vocabulary continuous speech recognition (LVCSR) systems. The goal is to improve the speech recognition accuracy of such a recognition system in a domain that it was not trained for, using untranscribed adaptation data.

This problem is particularly pressing as vast amounts of speech data are now available on the internet, from an ever-increasing range of domains and genres, such as broadcast shows, parliamentary proceedings, and other forms of public discourse and meetings. All such data should be made searchable and otherwise amenable to text processing, yet well-matched automatic speech recognition (ASR) systems are typically only available for a small subset of domains. In this paper, we apply a recognition system trained for the meeting recognition task evaluated by NIST [1] to congressional committee recordings available from the House Armed Services Committee (HASC) web site.

Previous studies on domain adaptation include supervised adaptation of acoustic models with discriminative criteria [2, 3], or unsupervised adaptation with maximum likelihood criterion [4]. Most work investigates adapting a single set of acoustic models. In this study, due to the fact that we don't have manual transcripts available for the HASC domain data except for the evaluation subset, we focus on unsupervised adaptation of acoustic models for a complex multipass LVCSR system with multiple sets of acoustic models and internal cross-adaptation stages. We investigate whether discriminative training criteria can be effective in unsupervised adaptation, and how different adaptation data selection and adaptation strategies affect the accuracy of intermediate and final recognition hypotheses.

The rest of this paper is organized as follows. In Section 2, the baseline LVCSR system is introduced. In Section 3, we describe the task and data used in this study. In Section 4, different adaptation data selection methods and adaptation strategies are discussed. In Section 5, we compare maximum likelihood and discriminative adaptation techniques. Section 6 presents experimental results, and Section 7 offers conclusions.

2. BASELINE MULTIPASS LVCSR SYSTEM

The baseline multipass LVCSR system is the SRI-ICSI meeting and lecture recognition system [5], as used in the NIST RT-07 evaluations [1]. The decoding architecture is depicted in Figure 1. Both Mel-frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP) front-ends are used. The decoding outputs of different acoustic models are combined twice using word confusion networks (denoted by crossed ovals in the figure). Except for the initial decodings, the acoustic models are cross-adapted to the output of a previous step using maximum likelihood linear regression (MLLR). Lattices are regenerated initially to speed up subsequent decoding steps. The lattices are regenerated once later to improve their accuracy, after adapting to the outputs of the first combination step. The lattice generation steps use within-word (nonCW) triphone models to generate high density lattices (thick lattices), and decoding from lattices uses two crossword (CW) models as MFCC-CW and PLP-CW. Each decoding step generates either lattices or N-best lists, both of which are rescored with a 4-gram language model (LM) that interpolates probability estimates from various source-specific LMs (meeting transcripts, telephone

[†]This work was performed while the author was at SRI.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
DEC 2010		**00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Unsupervised Domain Adaptation with Multiple Acoustic Models	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **SRI International,333 Ravenswood Avenue,Menlo Park,CA,94025**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We investigate the problem of adapting a recognition system with multiple acoustic models to a new domain in unsupervised mode. We compare maximum likelihood and discriminative approaches for unsupervised domain adaptation. Different adaptation data selection methods and adaptation strategies are investigated, using a baseline meeting recognition systemand adaptation data from a congressional committee web site. Experiments show that one should avoid adapting all acoustic models to the same recognition output, and that ASR confidence estimates improve results when used for rejecting low-quality ASR output. The results show 8% relative overall improvement from unsupervised adaptation.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**6**	

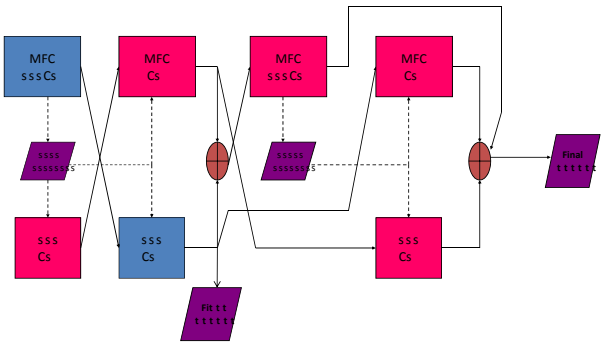

Fig. 1. Baseline multipass recognition system

conversations, web data, and broadcast news; see details in [6, 5]). N-best output is also rescored with duration models for phones and pauses. The rescored N-best lists from the three component systems, namely, MFCC-CW, MFCC-nonCW, and PLP-CW, are combined using N-best ROVER algorithm [7] and the final output is generated from this three-way system combination. The entire system runs around 3 times real time (3xRT) on Intel 3.0 GHz 2x4-core CPUs. More details about the baseline system are available in [5].

3. TASK

The task is to apply our baseline meeting recognition system to the HASC data domain and improve the recognition performance. The HASC video data is downloaded from the government web site at `http://armedservices.house.gov/hearing_information.shtml`. We downloaded 61 hours of videos from June 2009 to January 2010 and extracted the audio data with the `ffmpeg` package. The 61 hours of speech data are divided into a 50-hour adaptation set and an 11-hour evaluation set. Only the 11-hour evaluation set is manually transcribed and the 50-hour adaptation set has no manual transcription.

The baseline system acoustic models are trained on about 200 hours of meeting data, as well as two much larger corpora of non-meeting data. The MFCC-based models use about 1400 hours of telephone conversation as additional training data, whereas the PLP models are based on 900 hours of broadcast news as additional data; for details see [5]. All of these data sources are mismatched in style, topic, and acoustic conditions to the HASC target domain. Still, the final output of the baseline system has a word error rate (WER) of 17.5% on the evaluation set.

For adaptation purposes, the full baseline system is run on all 61 hours of audio data (including both adaptation set and evaluation set); the resulting hypotheses are used as adaptation data for the unsupervised adaptation experiments reported here.

4. ADAPTATION STRATEGIES

To adapt the acoustic models to a new domain, MLLR [8] and maximum a posteriori probability (MAP) [9] methods can be used. While it has been shown that MLLR can be effective with a limited amount of adaptation data, MAP is typically more effective with larger amounts of in-domain data. If transcriptions of the in-domain data are available, *supervised* MLLR or MAP adaptation can be performed. Without transcriptions, recognition outputs from the baseline system need to be used, giving *unsupervised* adaptation. In this paper, we focus on the unsupervised case, both because that is the case of greatest practical relevance, and because only a small amount of data (the evaluation portion) could be transcribed for this study.

Given the amount of adaptation data, we focus our investigation on MAP-adaptation only, and investigate alternative schemes to get the best use out of the available untranscribed adaptation data. At the same time, we hold constant the unsupervised MLLR adaptation strategy that is depicted in Figure 1. To clarify the distinction between the two levels of adaptation, system-internal MLLR adaptation uses only the test data being recognized, one HASC session at a time, whereas MAP adaptation uses all the pooled adaptation data, prior to recognition.

4.1. Adaptation Data Selection

Since all adaptation is unsupervised, the 11-hour test set of in-domain data can either be included in, or excluded from the unsupervised training set for MAP-adaptation. If we include the testing data and adapt on all 61 hours, there is more data for unsupervised MAP-adaptation. However, in this way, the acoustic models have already seen the test set and MLLR on the test data will be less effective. Furthermore, all three acoustic models will be rendered more similar, possibly reducing the benefits of system combination inside the recognition system. Therefore, we will investigate both including and excluding the test data from the MAP adaptation set.

4.2. Adaptation Topology

Another important issue is *adaptation topology*, that is, how to choose the unsupervised adaptation hypotheses for each set of acoustic models in MAP-adaptation. We investigate three different adaptation strategies: joint adaptation, self-adaptation, and extended cross-adaptation. The schemes are depicted graphically in Figure 4.2.

Joint adaptation

In joint adaptation, we use the final output hypothesis from the baseline system on the in-domain data (61 hours or 50 hours) to adapt all three sets of acoustic models: MFCC-nonCW, MFCC-CW, and PLP-CW models. Since in this case the same, best hypothesis is used for adapting all models, we expect the best decoding performance from each component

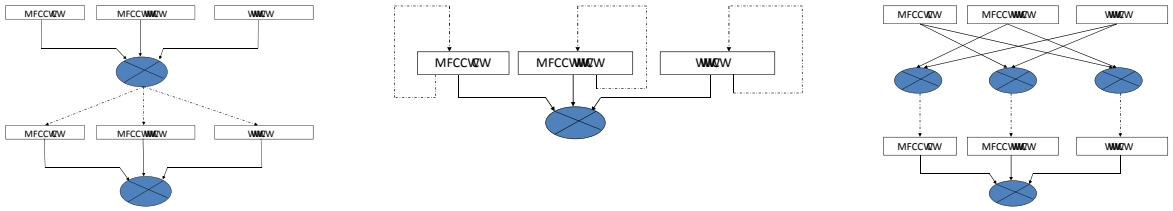

Fig. 2. Three adaptation topologies for acoustic model combination: joint adaptation (left), self adaptation (middle), and extended cross-adaptation (right).

system. However, since the final output of the adapted system is itself obtained by combining three (adapted) component systems, adapting all acoustic models to the same hypotheses may reduce their diversity and may not be the best choice after system combination in testing.

Self-adaptation

To make system combination more effective, one adaptation strategy is to adapt different acoustic models with different hypotheses. Self-adaptation is one choice: the output hypotheses of an acoustic model before system combination are used to adapt the corresponding acoustic model itself. For example, we only use the final 1-best decoding hypotheses (on 61 hours or 50 hours) from the MFCC-CW system to adapt the MFCC-CW model. In this way, during decoding, the component systems will be less correlated and more complementary in the information they contribute to system combination.

Extended cross-adaptation

The cross-adaptation topology adapts each model with the hypotheses from a different model, or in our extended implementation, the combined hypotheses from the rest of the models. The rationale of extended cross-adaptation is that one recognition system may be able to recognize some samples that are difficult for other systems. Therefore, each recognition system may be able to provide samples that are very informative for the other systems during adaptation. Also, compared to self-adaptation, extended cross-adaptation reduces the tendency of learning self-produced errors in the adaptation hypotheses. In the extended cross-adaptation topology for MAP-adaptation, we use the N-best-ROVER output of the MFCC-CW and MFCC-nonCW systems to adapt the PLP-CW model, the N-best ROVER output of the MFCC-nonCW and PLP-CW systems to adapt the MFCC-CW model, and the N-best ROVER output of the MFCC-CW and PLP-CW systems to adapt the MFCC-nonCW model. In this approach, diversity may still be preserved while there is less concern about accumulating errors when adapting to the output of a model itself. We denote the systems providing adaptation hypotheses the *teacher*, and the system using the adaptation hypotheses the *student*.

5. UNSUPERVISED DISCRIMINATIVE ADAPTATION

We also experiment with unsupervised MAP-adaptation using both maximum likelihood (ML) MAP and discriminative MAP.

5.1. ML-MAP Adaptation

Let us denote the unadapted mean and variance $\{\tilde{\mu}_{jm}, \tilde{\sigma}^2_{jm}\}$ for the m-th Gaussian in the j-th state. The adaptation data is denoted as $\mathcal{O} = \{o_1, \ldots, o_T\}$. The adapted model parameters can be effectively estimated by using count smoothing. The ML-MAP estimates for the adapted mean and variance are

$$\mu_{jm}^{\text{(ml-map)}} = \frac{\theta_{jm}(\mathcal{O}) + \tau\tilde{\mu}_{jm}}{\gamma_{jm} + \tau} \tag{1}$$

$$\sigma_{jm}^{\text{(ml-map)2}} = \frac{\theta_{jm}(\mathcal{O}^2) + \tau(\tilde{\mu}^2_{jm} + \tilde{\sigma}^2_{jm})}{\gamma_{jm} + \tau} - \mu_{jm}^{\text{(ml-map)2}} \tag{2}$$

where $\gamma_{jm} = \Sigma_{t=1}^{T}\gamma_{jm}(t)$ is the accumulated posterior probability of being in Gaussian m of state j, $\theta_{jm}(\mathcal{O}) = \Sigma_{t=1}^{T}\gamma_{jm}(t)o_t$, $\theta_{jm}(\mathcal{O}^2) = \Sigma_{t=1}^{T}\gamma_{jm}(t)o_t^2$, and τ is the smoothing factor. The larger τ is, the closer the update will be to the unadapted parameters.

5.2. Discriminative MAP Adaptation

Standard ML-MAP has been extended to incorporate discriminative training criteria such as MMI and MPE [10]. Discriminative MAP has two steps of operation. In the first step, the unadapted mean and variance are adapted with ML-MAP. In the second step, discriminative training is performed and the ML-MAP updated parameters are used as the prior for I-smoothing. The count weighting for this prior is set using an additional smoothing variable τ^I. For example, the MMI-MAP mean is given by

$$\mu_{jm}^{\text{(mmi-map)}} = \frac{\{\theta_{jm}^{\text{num}}(\mathcal{O}) - \theta_{jm}^{\text{den}}(\mathcal{O})\} + D_{jm}\hat{\mu}_{jm} + \tau^I\mu_{jm}^{\text{(ml-map)}}}{\{\gamma_{jm}^{\text{num}} - \gamma_{jm}^{\text{den}}\} + D_{jm} + \tau^I} \tag{3}$$

Table 1. *WER (%) results on the 11-hour evaluation set with different adaptation data sets. The 61hrs-set includes the test data, while the 50hrs-set excludes it.*

	First output	Final output
Baseline	31.0	17.5
ML-MAP on 50hrs	27.1	16.5
ML-MAP on 61hrs	24.2	17.1

where γ_{jm}^{num} and $\theta_{jm}^{\text{num}}(\mathcal{O})$ are the numerator statistics, γ_{jm}^{den} and $\theta_{jm}^{\text{den}}(\mathcal{O})$ are the denominator statistics, $\hat{\mu}_{jm}$ is the model mean for the previous iteration of the MMI training, and D_{jm} is the Gaussian-dependent parameter for the extended Baum-Welch (EBW) algorithm.

MMI-MAP has been successfully applied in supervised discriminative adaptation of a single set of acoustic models [2]. In this work, we will evaluate its efficacy on unsupervised adaptation for a multipass system. Furthermore, we will investigate the use of boosted MMI (BMMI) [11] as the discriminative training criterion in MMI-MAP. BMMI is a variant of MMI whereby the likelihoods of paths in the denominator lattice that have a relatively higher phone error are boosted.

6. EXPERIMENTAL RESULTS

6.1. Adaptation Strategies

First, we evaluate whether it is beneficial to include the 11-hour evaluation set in the unsupervised MAP-adaptation. ML-MAP experiments are performed with the full 61-hour adaptation set and the 50-hour adaptation set. The ML-MAP smoothing factor τ is set to 20. All three acoustic models are adapted with the final baseline output, that is, the **joint adaptation** topology is used. The adapted models are then used to re-run decoding of the evaluation set. For evaluation, we look at both the initial decoding results from the MFCC nonCW stage and the final system output. The results are shown in Table 1. By including the same 11-hour evaluation set in the unsupervised adaptation, the initial decoding results are 2.9% absolute better than with the 50-hour adaptation set. However, the final performance is 0.6% absolute worse with the 61-hour adaptation set. This observation is consistent with our hypothesis that including the evaluation data in unsupervised MAP-adaptation dramatically reduces diversity between the three component systems and left little room for improvement from MLLR cross-adaptation and system combination in decoding. Therefore, to improve the final performance of the multipass LVCSR system, the evaluation data should be excluded from the unsupervised adaptation data in MAP-adaptation.

Next, we compare different adaptation topologies for MAP-adaptation using the 50-hour adaptation set. The results are shown in Table 2. The results from each of the three subsystems before N-best ROVER combination are also compared. Although self-adaptation has worse results in all three subsystems before combination, it gives similar results to the joint adaptation after combination, presumably because the subsystems are more complementary. Extended cross-adaptation gives slightly better results over both joint adaptation and self-adaptation, apparently striking the best balance between subsystem improvement and complementarity.

Table 2. *WER (%) results with different MAP-adaptation topologies.*

	mel-cw	mel-noncw	plp-cw	final
Baseline	19.1	20.9	18.0	17.5
Joint adaptation	17.9	19.3	16.8	16.5
Self-adaptation	18.3	19.4	17.2	16.6
Extended Cross-adaptation	18.0	19.3	16.9	16.3

Table 3. *WER (%) results with ML-MAP and BMMI-MAP adaptations.*

	mel-cw	mel-noncw	plp-cw	final
Baseline	19.1	20.9	18.0	17.5
ML-MAP	18.0	19.3	16.9	16.3
BMMI-MAP	17.9	19.4	17.4	16.5

6.2. Discriminative Adaptation

Several discriminative adaptation techniques are examined in this unsupervised adaptation task. We adopt the extended cross-adaptation topology for the following experiments. The I-smoothing factor τ^I is set to 25 in BMMI-MAP. As shown in Table 3, BMMI-MAP performed similar or slightly worse than ML-MAP. In addition, as shown in Table 4, BMMI-MAP performance tends to drop quickly over iterations. This indicates that the errors in the adaptation hypothesis are more detrimental to discriminative adaptation. This observation is consistent with the findings on unsupervised discriminative acoustic model training. Since discriminative training aims to reduce the difference between the recognized output and the correct transcription, where in the case of unsupervised training/adaptation the "correct" transcriptions are in fact errorful recognition output, discriminative training/adaptation is therefore far more sensitive to the accuracy of the transcriptions than ML training [12].

6.3. Data Selection/Filtering

In the previous experiment, we have shown that extended cross-adaptation results are just slightly better than self-adaptation and joint adaptation. In that experiment, we used all recognition hypotheses from other subsystems to adapt the

Table 4. *WER (%) results with BMMI-MAP adaptations over iterations.*

		mel-cw	mel-noncw	plp-cw
Baseline		19.1	20.9	18.0
ML-MAP		18.0	19.3	16.9
BMMI-MAP	iter1	17.9	19.4	17.4
	iter2	18.4	19.7	17.7
	iter3	18.7	19.4	17.8

current model. We denote this strategy the *naive* approach. However, these adaptation hypotheses are quite noisy (word error rate in the range of 18% to 20%). To avoid using very errorful data for MAP-adaptation, we developed a data selection/filtering approach, denoted *max-conf*. In this approach, we select adaptation hypotheses that have confidence scores above a certain threshold θ from the *teacher* system.

We approximated the confidence score of a hypothesis with a weighted sum of its word confidence scores.

$$\text{Conf}(h) = \frac{\sum_{i=1}^{N} \text{Conf}(w_i) \cdot \text{Dur}(w_i)}{\sum_{i=1}^{N} \text{Dur}(w_i)} \quad (4)$$

where N is the number of words in the hypothesis h, and $\text{Conf}(w_i)$ and $\text{Dur}(w_i)$ are the confidence score and duration of the word w_i, respectively. The confidence score thus computed is an estimate for the per-frame average probability of correct recognition.

To compute the word-level confidence scores, we used a neural network that takes several word-level features as input [13]. The main input feature used is the word posterior probability obtained from the N-best ROVER algorithm. Additional minor features include the overall length of the hypothesis and the normalized relative position of the word in the word string. The neural network was trained on an English broadcast news test set. System output had been labeled as correct or incorrect by a dynamic alignment between the hypothesis and the reference word strings. Two output nodes were used (2 classes: correct and incorrect), with softmax output layers. The training criterion was cross-entropy minimization. The network had one hidden layer with 4 hidden nodes.

We conducted a grid search to optimize the minimum-confidence θ for the *max-conf* data selection/filtering approach and results are shown in Table 5. Note that the result from the *naive* approach is the same as shown in Table 2. We obtained the best final WER of 16.1% from extended cross-adaptation with $\theta = 0.7$. This way, extended cross-adaptation produced 1.4% absolute gain in WER over the baseline, and 0.4% and 0.5% absolute gain over joint-adaptation and self-adaptation, respectively.

In the previous discussion, we hypothesized that recognition errors could be more detrimental to discriminative adaptation than to ML adaptation. Hence, we also examined the effect of applying *max-conf* data selection when comparing ML-MAP and BMMI-MAP. The results are shown in Table 6. As can be seen, using *max-conf* data selection improved both ML-MAP and BMMI-MAP adaptation performance. In particular, with filtering adaptation data with $\theta = 0.85$, *max-conf* improved BMMI-MAP by 0.4% absolute over the *naive* approach. Unfortunately, even with *max-conf* data selection, BMMI-MAP is still not giving a gain over ML-MAP, both reaching 16.1% final WER.

7. CONCLUSIONS AND FUTURE WORK

We have presented an experimental study in unsupervised domain adaptation for a multipass LVCSR system with multiple acoustic models. We observed that it is better to exclude the evaluation data from the unsupervised adaptation. Adapting each acoustic model to a different intermediate hypothesis yields better results than adapting to the same combined hypothesis. Finally, unsupervised discriminative adaptation shows performance improvements similar to ML adaptation, after investigations of data selection/filtering approaches.

Future work can go in several directions. First, we plan to investigate other confidence score estimation approaches such as explicit sentence level confidence score estimation, conducting the sentence level data selection directly on the phone accuracy domain for discriminative training, and the use of state confidence scores for data selection. Second, we will investigate unsupervised discriminative adaptation with active learning, similar to the work of employing direct manual transcription for discriminative training [14]. Finally, we may incorporate unsupervised language model adaptation [15] in the framework presented here.

8. ACKNOWLEDGMENTS

The authors thank SRI colleague Jing Zheng for useful discussions. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. FA8750-09-D-0183 (approved for public release, distribution unlimited). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

9. REFERENCES

[1] J. G. Fiscus, J. Ajot, and J. S. Garofolo, "The Rich Transcription 2007 meeting recognition evaluation", in R. Stiefelhagen, R. Bowers, and J. Fiscus, editors, *Multimodal Technologies for Perception of Humans. International Evaluation Workshops CLEAR 2007 and RT 2007*, vol. 4625 of *Lecture Notes in Computer Science*, pp. 373–389. Springer, Berlin, 2008.

Table 5. *WER (%) results with different adaptation topologies, using* naive *and* max-conf *data selection approaches.*

		mel-cw	mel-noncw	plp-cw	final
Baseline		19.1	20.9	18.0	17.5
Extended	naive	18.0	19.3	16.9	16.3
Cross-adaptation	max-conf (θ=0.7)	17.8	19.0	16.7	16.1

Table 6. *WER (%) results with ML-MAP and BMMI-MAP adaptations, using* naive *and* max-conf *data selection approaches.*

		mel-cw	mel-noncw	plp-cw	final
Baseline		19.1	20.9	18.0	17.5
ML-MAP	naive	18.0	19.3	16.9	16.3
	max-conf (θ=0.7)	17.8	19.0	16.7	16.1
BMMI-MAP	naive	17.9	19.4	17.4	16.5
	max-conf (θ=0.85)	17.5	19.1	17.0	16.1

[2] M. Gales, Y. Dong, D. Povey, and P. Woodland, "Porting: Switchboard to the voicemail task", *in Proc. ICASSP*, pp. 536–539, 2003.

[3] J. Zheng and A. Stolcke, "fMPE-MAP: Improved discriminative adaptation for modeling new domains", *in Proc. Interspeech*, pp. 1573–1576, Antwerp, Aug. 2007.

[4] D. Giuliani and M. Federico, "Unsupervised language and acoustic model adaptation for cross domain portability", *in Proc. ISCA ITR Workshop, Sophia-Antipolis*, 2001.

[5] A. Stolcke, K. Boakye, Özgür Çetin, A. Janin, M. Magimai-Doss, C. Wooters, and J. Zheng, "The SRI-ICSI Spring 2007 meeting and lecture recognition system", in R. Stiefelhagen, R. Bowers, and J. Fiscus, editors, *Multimodal Technologies for Perception of Humans. International Evaluation Workshops CLEAR 2007 and RT 2007*, vol. 4625 of *Lecture Notes in Computer Science*, pp. 450–463, Berlin, 2008. Springer.

[6] Ö. Çetin and A. Stolcke, "Language modeling in the ICSI-SRI Spring 2005 meeting speech recognition evaluation system", Technical Report TR-05-06, International Computer Science Institute, Berkeley, CA, 2005.

[7] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauché, C. Richey, E. Shriberg, K. Sönmez, F. Weng, and J. Zheng, "The SRI March 2000 Hub-5 conversational speech transcription system", *in Proceedings NIST Speech Transcription Workshop*, College Park, MD, May 2000.

[8] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition", *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[9] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 291–299, 1994.

[10] D. Povey, M. Gales, D. Kim, and P. Woodland, "MMI-MAP and MPE-MAP for acoustic model adaptation", *in Proc. Eurospeech*, pp. 1981–1984, 2003.

[11] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training", *in Proc. ICASSP*, pp. 4057–4060, 2008.

[12] L. Wang, M. F. J. Gales, and P. C. Woodland, "Unsupervised training for Mandarin broadcast news and conversation transcription", *in Proc. ICASSP*, vol. 4, pp. 353–356, Honolulu, Apr. 2007.

[13] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig, and A. Stolcke, "Neural-network based measures of confidence for word recognition", *in Proc. ICASSP*, vol. 2, pp. 887–890, Munich, Apr. 1997.

[14] K. Yu, M. J. F. Gales, and P. C. Woodland, "Unsupervised training with directed manual transcription for recognizing Mandarin broadcast audio", *in Proc. Interspeech*, pp. 1709–1712, Antwerp, Aug. 2007.

[15] G. Tur and A. Stolcke, "Unsupervised language model adaptation for meeting recognition", *in Proc. ICASSP*, vol. 4, pp. 173–176, Honolulu, Apr. 2007.